ISyE 6416: Computational Statistics
Homework 3
(25 points for each question. 100 points total.)

This homework is due on **Feb. 21, 2019**.

• Please write your team member's name is you collaborate.

1. **Deriving M-step in EM for GMM.** (15 points)

   (a) Derive forms of π_c, μ_c and Σ_c that maximize the $Q(\theta|\theta_k)$ function:

$$Q(\theta|\theta_k) = \sum_{i=1}^{n}\sum_{c=1}^{C} p_{i,c} \log \pi_c + \sum_{i=1}^{n}\sum_{c=1}^{C} p_{i,c} \log \phi(x_i|\mu_c, \Sigma_c)$$

   where $\theta = [\pi_c, \mu_c, \Sigma_c]$ subject to the constraint $\sum_{c=1}^{C} \pi_c = 1$, are given by

$$\pi_c^{(k+1)} = \frac{\sum_{i=1}^{n} p_{i,c}}{n}$$

$$\mu_c^{(k+1)} = \frac{\sum_{i=1}^{n} p_{i,c} x_i}{\sum_{i=1}^{c} p_{i,c}}$$

$$\Sigma_c^{(k+1)} = \frac{\sum_{i=1}^{n} p_{i,c}(x_i - \mu_c)(x_i - \mu_c)^T}{\sum_{i=1}^{n} p_{i,c}}$$

$$p_{i,c} \propto \pi_c^{(k)} \phi(x_i|\mu_c^{(k)}, \Sigma_c^{(k)})$$

2. **Implementing EM algorithm for MNIST dataset.** (25 points)

   Implement the EM algorithm for fitting a Gaussian mixture model for the MNIST dataset. We reduce the dataset to be only two cases, of digits "2" and "6" only. Thus, you will fit GMM with $C = 2$. Use the data file data.mat or data.dat on Canvas. True label of the data are also provided in label.mat and label.dat

   The matrix images is of size 784-by-1990, i.e., there are totally 1990 images, and each column of the matrix corresponds to one image of size 28-by-28 pixels (the image is vectorized; the original image can be recovered, e.g., using MATLAB code, reshape(images(:,1),28, 28).

   (a) Select from data one raw image of "2" and "6" and visualize them, respectively.

   (b) Use random Gaussian vector with zero mean as initial means, and identity matrix as initial covariance matrix for the three clusters. Please plot the log-likelihood function versus the number of iterations to show your algorithm is converging.

   (c) Report the finally fitting GMM model when EM terminates: the weights for each component, the mean vectors (please reformat the vectors into 28-by-28 images and show these images in your submission). Ideally, you should be able to see these means corresponds to "average" images. No need to report the covariance matrices.

(d) (Optional). Use the p_{ic} to infer the labels of the images, and compare with the true labels. Report the miss classification rate.

3. **EM application - fair review system.** (25 points)

Consider the following problem. There are P papers submitted to a machine learning conference. Each of R reviewers reads each paper, and gives it a score indicating how good he/she thought that paper was. We let $x^{(pr)}$ denote the score that reviewer r gave to paper p. A high score means the reviewer liked the paper, and represents a recommendation from that reviewer that it be accepted for the conference. A low score means the reviewer did not like the paper.

We imagine that each paper has some "intrinsic" true value that we denote by μ_p, where a large value means it's a good paper. Each reviewer is trying to estimate, based on reading the paper, what μ_p is; the score reported $x^{(pr)}$ is then reviewer rs guess of μ_p.

However, some reviewers are just generally inclined to think all papers are good and tend to give all papers high scores; other reviewers may be particularly nasty and tend to give low scores to everything. (Similarly, different reviewers may have different amounts of variance in the way they review papers, making some reviewers more consistent/reliable than others.) We let ν_r denote the "bias" of reviewer r. A reviewer with bias ν_r is one whose scores generally tend to be ν_r higher than they should be.

All sorts of different random factors influence the reviewing process, and hence we will use a model that incorporates several sources of noise. Specifically, we assume that reviewers's scores are generated by a random process given as follows:

$$y^{(pr)} \sim \mathcal{N}(\mu_p, \sigma_p^2)$$

$$z^{(pr)} \sim \mathcal{N}(\nu_r, \tau_r^2)$$

$$x^{(pr)} | y^{(pr)}, z^{(pr)} \sim \mathcal{N}(y^{(pr)} + z^{(pr)}, \sigma^2).$$

The variables $y^{(pr)}$ and $z^{(pr)}$ are independent; the variables (x, y, z) for different paper-reviewer pairs are also jointly independent. Also, we only ever observe the $x^{(pr)}$s; thus, the $y^{(pr)}$s and $z^{(pr)}$s are all latent random variables.

We would like to estimate the parameters μ_p, σ_p^2, ν_r, τ_r^2. If we obtain good estimates of the papers "intrinsic values" μ_p, these can then be used to make acceptance/rejection decisions for the conference.

We will estimate the parameters by maximizing the marginal likelihood of the data $\{x^{(pr)}; p = 1, \ldots, P, r = 1, \ldots, R\}$. This problem has latent variables $y^{(pr)}$s and $z^{(pr)}$s, and the maximum likelihood problem cannot be solved in closed form. So, we will use EM. Your task is to derive the EM update equations. For simplicity, you need to treat only $\{\mu_p, \sigma_p^2; p = 1 \ldots, P\}$ and $\{\nu_r, \tau_r^2; r = 1...R\}$ as parameters. I.e. treat σ^2 (the conditional variance of $x^{(pr)}$ given $y^{(pr)}$ and $z^{(pr)}$) as a fixed, known constant.

(a) In this part, we will derive the E-step:

i. The joint distribution $p(y^{(pr)}, z^{(pr)}, x^{(pr)})$ has the form of a multivariate Gaussian density. Find its associated mean vector and covariance matrix in terms of the parameters μ_p, σ_p^2, ν_r, τ_r^2 and σ^2. [Hint: Recognize that $x^{(pr)}$ can be written as $x^{(pr)} = y^{(pr)} + z^{(pr)} + \epsilon^{(pr)}$, where $\epsilon^{(pr)} \sim \mathcal{N}(0, \sigma^2)$ is independent Gaussian noise.

ii. Derive an expression for $Q_{pr}(\theta'|\theta) = \mathbb{E}[\log p(y^{(pr)}, z^{(pr)}, x^{(pr)})|x^{(pr)}, \theta]$ using the conditional distribution $p(y^{(pr)}, z^{(pr)}|x^{(pr)})$ (E-step) (Hint, you may use the rules for conditioning on subsets of jointly Gaussian random variables.)

(b) Derive the M-step updates to the parameters μ_p, σ_p^2, ν_r, and τ_r^2. [Hint: It may help to express an approximation to the likelihood in terms of an expectation with respect to $(y^{(pr)}, z^{(pr)})$ drawn from a distribution with density $Q_{pr}(y^{(pr)}, z^{(pr)})$.]

Remark: In a recent machine learning conference, John Platt (whose SMO algorithm you've seen) implemented a method quite similar to this one to estimate the papers' true scores. (There, the problem was a bit more complicated because not all reviewers reviewed every paper, but the essential ideas are the same.) Because the model tried to estimate and correct for reviewers' biases, its estimates of the paper's value were significantly more useful for making accept/reject decisions than the reviewers? raw scores for a paper.

4. **Proof (15 points).** Prove the following results in the lecture slides for hidden Markov model:

$$\mathbb{P}(S_t = i|o_1, \cdots, o_T) \propto \underbrace{\mathbb{P}(S_t = i, o_1, \cdots, o_t)}_{\alpha_i(t)} \cdot \underbrace{\mathbb{P}(o_{t+1}, \cdots, o_T|S_t = i)}_{\beta_i(t)}$$

$$\mathbb{P}(S_t = i, S_{t+1} = j|o_1, \ldots, o_T) \propto \alpha_i(t)\beta_j(t+1)a_{i,j}b_{j,o_{t+1}}$$

5. **The occasionally dishonest casino (20 points).** Consider the occasionally dishonest casino problem we discussed in class. There are two states for the dice, the "fair" (F) and "loaded" (L) dices, and under each state, the emission probability matrix is given by

$$E = \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/2 \end{bmatrix}$$

And the state transition matrix is

$$P(F|F) = 0.95, \quad P(L|F) = 0.05, \quad P(L|L) = 0.9, \quad P(H|L) = 0.1.$$

Assume the observed sequence is 445436316566265666. Assume the initial distribution to be [1/2, 1/2].

(a) Program your own code to answer this question: use the forward-backward algorithm to determine the probability of being in each state at time k given all three observed bits.

(b) What is the most likely sequence of states given all the observations? Implement the computation.